데이터 관련 용어 정리

## Data

**정형데이터** - Machine Readable, 스키마 형식에 맞게 저장된 데이터

**비정형데이터** - 구조가 일정하지 않은 데이터, 관계형모델에 적합하지 않은 데이터, 규격화된 데이터필드에 저장되지 않은 데이터

**반정형데이터** - 정형인데 Meta Data가 아님. 정형구조의 데이터 모델을 준수하지 않음. ex) 센서값, 웹로그값

**Metadata** - 데이터에 대한 데이터, Karen Coyle에 의하면 "어떤 목적을 가지고 만들어진 데이터(Constructed data with a purpose)"라고도 정의한다. 도서관에서 책을 찾을 때를 생각하면 쉽다(도서관의 서지기술용). Machine Actionable 한 형태의 메타데이터가 많이 사용되고 있다. 또 예를들어 1,2,3,4,5∙∙∙ 처럼 일련의 넘버가 DB에 저장이 되어있을 때 단순히 1,2,3,4,5 들은 데이터지만 이 데이터에 대한 의미를 알게해주는 넘버 는 메타데이터가 된다.

**Big Data** - 기존 데이터베이스 관리 도구로 데이터를 수집, 저장, 관리, 분석할 수 있는 역량을 넘어서는 대량의 정형 또는 비정형 데이터 집합 및 이러한 데이터로부터 가치를 추출하고 결과를 분석하는 기술.  
분석 기술로는 데이터마이닝, 기계 학습(머신 러닝), 자연 언어 처리, 패턴 인식 등이 해당된다. 분석 인프라로는 [하둡](#Hadoop)이 있고 더 유연하고 빠르게 처리하기 위해 NoSQL 기술이 활용되기도 한다.  
대표적인 표현기술로는 R이 있다.

#### Data Profiling – 비정형 데이터와 반정형 데이터를 분석하기 위해 정형화하고 Metadata를 정의하는 과정을 Data Profiling 이라고 한다.

**Data Warehouse와 Data Mart** – DW는 데이터를 찾기 쉽게 저장해놓은 거대한 창고라면, DM은 데이터를 보기 쉽게 진열한 마트와 같다.

**Data Governance** **–** 데이터 거버넌스는 기업에서 사용하는 데이터의 가용성, 유용성, 통합성, 보안성을 관리하기 위한 정책과 프로세스를 다루며 프라이버시, 보안성, 데이터 품질, 관리규정 준수를 강조한다.

**PPM –** Process Performance Management, 각 프로세스의 효율성, 적시성 등을 평가하는 것이 PPM임. 주된 평가 관점은 신뢰성, 책임성, 유연성, 비용, 자산 이다.

**BAM –** Business Activity Monitoring, 프로세스에 대한 실시간 감시와 조절. 예를 들어 비행기가 10분 연착된다는 정보를 받았을 때, 손실을 최소화하기 위해 즉각적으로 그에 대한 피드백을 하는 것이 BAM이 될 수 있다.

## Hadoop

**Hadoop** – High-Availability Distributed Object- Oriented Platform의 약자로 대량의 자료를 처리할 수 있는 큰 컴퓨터 클러스터에서 동작하는 분산 응용 프로그램을 지원하는 “프리웨어 자바 소프트웨어 프레임워크”이다. 프로그래밍 언어는 자바이며, 더그 커팅이 자신의 아들이 놀던 코끼리 장난감의 이름을 따서 하둡이라고 지었다. 코끼리는 빅데이터를 상징하는 동물이 되었다. 하둡은 분산 프로그래밍에 대해서 IT업계의 요구를 충족시켜주는 Software인데, 분산저장(HDFS)와 분산 처리기술(Map Reduce)이 탁월 하다.

**HDFS** – Hadoop Distributed File System의 약자로 하둡 프레임워크를 위해 자바 언어로 작성된 분산 확장 파일 시스템이다. 여러 기계에 대용량 파일들을 나눠서 저장한다. 데이터들을 여러 서버에 중복해서 저장을 함으로써 데이터 안정성을 얻는다. 따라서 호스트에 RAID 저장장치를 사용하지 않아도 된다. 마스터/슬레이브(master/slave) 구조를 가진다.

* **하드웨어 오동작**: 하드웨어 수가 많아지면 그중에 일부 하드웨어가 오동작하는 것은 예외 상황이 아니라 항상 발생하는 일이다. 따라서 이런 상황에서 빨리 자동으로 복구하는 것은 HDFS의 중요한 목표다.
* **스트리밍 자료 접근**: 범용 파일 시스템과 달리 반응 속도보다는 시간당 처리량에 최적화되어 있다.
* **큰 자료 집합**: 한 파일이 기가바이트나 테라바이트 정도의 크기를 갖는 것을 목적으로 설계되었다. 자료 대역폭 총량이 높고, 하나의 클러스터에 수 백개의 노드를 둘 수 있다. 하나의 인스턴스에서 수천만여 파일을 지원한다.
* **간단한 결합 모델**: 한번 쓰고 여러번 읽는 모델에 적합한 구조이다. 파일이 한번 작성되고 닫히면 바뀔 필요가 없는 경우를 위한 것이다. 이렇게 함으로써 처리량을 극대화할 수 있다.
* **자료를 옮기는 것보다 계산 작업을 옮기는 것이 비용이 적게 든다**: 자료를 많이 옮기면 대역폭이 많이 들기 때문에 네트워크 혼잡으로 인하여 전체 처리량이 감소한다. 가까운 곳에 있는 자료를 처리하게 계산 작업을 옮기면 전체적인 처리량이 더 높아진다.
* **다른 종류의 하드웨어와 소프트웨어 플랫폼과의 호환성:** 서로 다른 하드웨어와 소프트웨어 플랫폼들을 묶어 놓아도 잘 동작한다.

**DBMS –** 다수의 사용자들이 데이터베이스 내의 데이터를 접근할 수 있도록 해주는 소프트웨어 도구의 집합이다. DBMS는 사용자 또는 다른 프로그램의 요구를 처리하고 적절히 응답하여 데이터를 사용할 수 있도록 해준다.

DBMS의 기능으로는 정의, 구축, 조작, 공유, 보호, 유지보수가 있다.

DBMS는 자료의 통합성을 증진시킨다(DBMS는 자료와의 관계성을 정의하기 때문에 자료 통합이 증진된다). 데이터의 접근성이 용이하다. 데이터 통제가 강화된다. Application 프로그램들을 쉽게 개발하고 관리할 수 있다. 보안이 강화된다.

**Data Mining** –대규모로 저장된 데이터 안에서 체계적이고 자동적으로 통계적 규칙이나 패턴을 찾아 내는 것이다. 다른 말로는 KDD(데이터베이스 속의 지식 발견, knowledge-discovery in databases)라고도 일컫는다. 다양한 응용 분야에서 기술적으로 많이 쓰이지만 **단점**이 하나 있는데, 자료에 의존하여 현상을 해석하고 개선하려고 하기 때문에 자료가 현실을 충분히 반영하지 못한 상태에서 정보를 추출한 모형을 개발할 겨우 잘못된 모형을 구축하는 오류를 범할 수가 있다는 것이다. 적용 분야는 다음과 같다.

* 군집화(Clustering): 구체적인 특성을 공유하는 군집을 찾는다. 군집화는 미리 정의된 특성에 대한 정보를 가지지 않는다는 점에서 분류와 다르다 (예: 유사 행동 집단의 구분)
* 분류(Classification): 일정한 집단에 대한 특정 정의를 통해 분류 및 구분을 추론한다 (예: 경쟁자에게로 이탈한 고객)
* 연관성(Association): 동시에 발생한 사건간의 관계를 정의한다. (예: 장바구니안의 동시에 들어 가는 상품들의 관계 규명)
* 연속성(Sequencing): 특정 기간에 걸쳐 발생하는 관계를 규명한다. 기간의 특성을 제외하면 연관성 분석과 유사하다 (예: 슈퍼마켓과 금융상품 사용에 대한 반복 방문)
* 예측(Forecasting): 대용량 데이터집합내의 패턴을 기반으로 미래를 예측한다 (예: 수요예측)

**Archive –** 보관하다, 모으다 라는 뜻으로 온라인 디스크에 있는 데이터를 보조기억장치로 이관해서 선별적인 데이터에 대한 보존과 저장, 데이터를 활용하기위한 동일한 온라인 접속을 허용하는 기술을 말한다. Storage와 구별해야하는 점은 이 녀석은 data에 대한 단순 backup에 의미를 두고 있고, Archive는 선별적 backup과 압축 보관, 복구 및 검색 등에 활용한다는 점이다.

**OLAP –** Online Analytical Processing의 약자로 말 그대로 ‘온라인 분석 처리’이다. 의사결정 지원 시스템의 대표적인 예로 직접적인 대화형 방식으로 다차원(보통 DB는 2차원이지만 현업에서는 다차원을 요구한다)의 데이터 분석을 할 수 있도록 도와준다. 1993년 에드거 F. 커드에 의해 처음 제안됐다.   
 특성으로는 다차원 정보 제공, 중간 매개자 없이 사용자가 직접 데이터 접근, 대화형태 정보분석, 의사 결정 지원이 있다. Data Mining과 다른 점은 OLAP는 사람이 직접 분석하는 거고, Mining은 분석 툴을 이용한다는 것이다.

**OLTP –** Online Transaction Processing의 약자로 application을 손쉽게 관리할 수 있도록 도와주는 정보 시스템의 한 계열로서, 일반적으로 데이터 기입 및 트랜잭션 처리를 위해 존재한다. 은행의 ATM이 상용 transaction 처리를 응용한 하나의 예로 볼 수 있다.

**EUC –** 정보를 필요로 하는 사람 스스로가 DB에 접근하여 자신이 원하는 정보를 얻고 입력하거나 수정할 수 있게 한 것을 말한다. BISS(Business Intelligence Self Service)와 같은 의미이며, Meta Data의 관리가 필수적으로 요구된다.

**MPP** – Massively Parallel Processing 분산 memory 병렬처리. 모든 processor들은 독립적인 memory를 가지기 때문에 memory에서의 access충돌이 일어나지 않고, processor의 독립성이 높기 때문에 processor를 늘리면 performance가 증가한다.

**SMP –** Symmetric Multi Processing. 공통의 memory를 이용연계하는 병렬처리법. 동작에 필요한 Data를 공통으로 가지고 있기 때문에 효율성이 높다. 하지만 Processor가 늘어나면 memory 순서 부여에 문제가 생기고 따라서 access충돌이 발생한다.

**Map Reduce(분산 처리 기술) –** 대용량 데이터를 분산 처리하기 위한 목적으로 개발된 프로그래밍 모델. Google에 의해 고안됐음. 병렬 처리 기법 중 하나로 주목을 받음. 맵 리듀스는 순서대로 정리된 데이터를 분산 처리(map)하고 이를 다시 합치(Reduce,)는 과정을 거친다.